

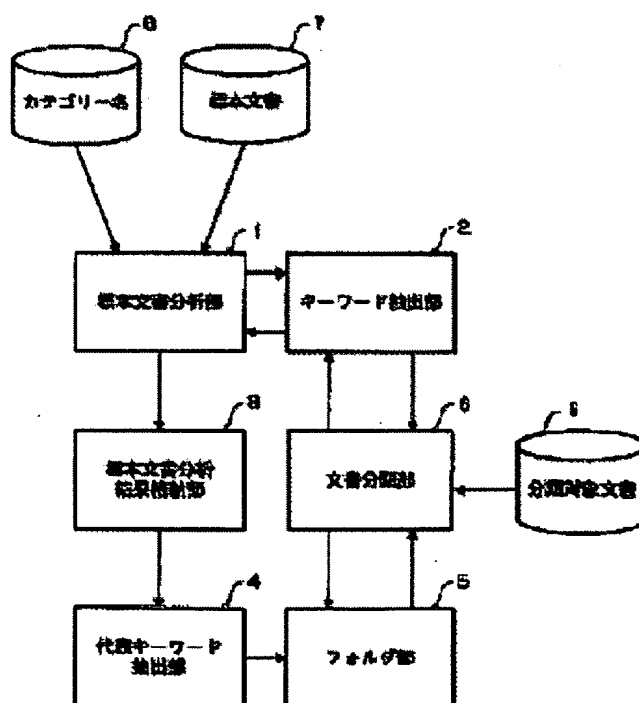
DOCUMENT SORTING DEVICE AND MACHINE-READABLE RECORDING MEDIUM RECORDING PROGRAM

Publication number: JP11025121
Publication date: 1999-01-29
Inventor: NAGATSUMA HIDEAKI
Applicant: NIPPON ELECTRIC CO
Classification:
 - international: G06F17/30; G06F17/30; (IPC1-7): G06F17/30
 - European:
Application number: JP19970189040 19970630
Priority number(s): JP19970189040 19970630

Report a data error here

Abstract of JP11025121

PROBLEM TO BE SOLVED: To prevent the occurrence of events disabling preparation of a sorting dictionary and to quickly prepare the sorting dictionary in a document sorting device for preparing a sorting dictionary and for sorting a document to be sorted in each category by using the dictionary. **SOLUTION:** A user inputs plural sample documents 7 and a category name 8 to which these documents 7 belong in each of plural categories to which documents are to be sorted. A sample document analysis part 1 extracts a keyword included in each sample document 7 by the use of a keyword extraction part 2 and stores a pair of the extracted keyword and its corresponding category name 8 in a storing part 3 to store a sample document analysis result. A representative keyword extraction part 4 extracts a keyword included in all sample documents 7 belonging to each category as a representative keyword based on the pair of the keyword and the category name stored in the storing part 3, and stores a pair of the extracted representative keyword and the category name in a holder part 5.



Data supplied from the esp@cenet database - Worldwide

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-25121

(43) 公開日 平成11年(1999) 1月29日

(51) Int.Cl.⁶

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/401

3 1 0 D

3 1 0 A

審査請求 有 請求項の数 5 F D (全 10 頁)

(21) 出願番号 特願平9-189040

(22) 出願日 平成9年(1997) 6月30日

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 長妻 秀明

東京都港区芝五丁目7番1号 日本電気株式会社内

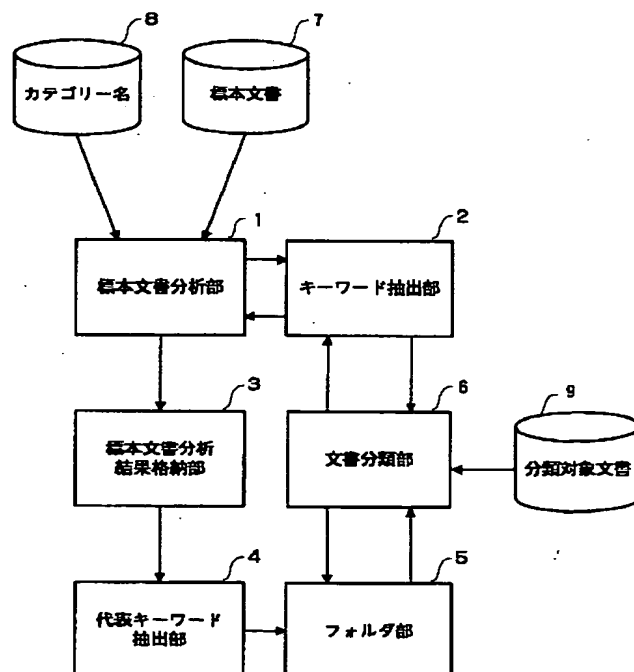
(74) 代理人 弁理士 境 廣巳

(54) 【発明の名称】 文書分類装置及びプログラムを記録した機械読み取り可能な記録媒体

(57) 【要約】

【課題】 分類用辞書を作成し、それを用いて分類対象文書をカテゴリー別に分類する文書分類装置に於いて、分類用辞書作成不能となる事態の発生を極力抑え、且つ分類用辞書を短時間で作成できるようにする。

【解決手段】 ユーザは、分類先となる複数のカテゴリーそれぞれについて、複数の標本文書7とそれが属するカテゴリー名8を入力する。標本文書分析部1は、キーワード抽出部2を用いて標本文書7に含まれているキーワードを抽出し、抽出したキーワードとカテゴリー名8とを対にして標本文書分析結果格納部3に格納する。代表キーワード抽出部4は、標本文書分析結果格納部3中のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書7に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にしてフォルダ部5に格納する。



(2)

【特許請求の範囲】

【請求項1】 分類用辞書と、

標本文書分析結果格納部と、
 文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部と、
 前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する代表キーワード抽出部と、
 前記分類用辞書を使用して分類対象文書をカテゴリー別に分類する文書分類部とを備えたことを特徴とする文書分類装置。

【請求項2】 前記代表キーワード抽出部の代わりに、前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書の内の、前記代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する拡張代表キーワード抽出部を備えたことを特徴とする請求項1記載の文書分類装置。

【請求項3】 前記文書分類部は、ユーザによって入力された分類対象文書に含まれるキーワードを抽出し、更に、前記分類用辞書を参照して前記抽出したキーワードと一致する代表キーワードを最も多く含むカテゴリーを前記分類対象文書の分類先とする構成を備えたことを特徴とする請求項1または2記載の文書分類装置。

【請求項4】 分類用辞書と標本文書分析結果格納部とを備えたコンピュータを、
 文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部、
 前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対に基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する代表キーワード抽出部、
 前記分類用辞書を使用して分類対象文書をカテゴリー別

に分類する文書分類部として機能させるためのプログラムを記録した機械読み取り可能な記録媒体。

【請求項5】 分類用辞書と標本文書分析結果格納部とを備えたコンピュータを、
 文書の分類先となる複数のカテゴリーそれぞれについて、そのカテゴリーに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリーのカテゴリー名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部、
 前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリー名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリー毎に、そのカテゴリーに属する全ての標本文書の内の、前記代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリー名とを対にして前記分類用辞書に格納する拡張代表キーワード抽出部、

前記分類用辞書を使用して分類対象文書をカテゴリー別に分類する文書分類部として機能させるためのプログラムを記録した機械読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、電子化された文書をカテゴリー別に自動的に分類する技術に関する。

【0002】

【従来の技術】ワードプロセッサ等によって作成された電子化された文書をカテゴリー別に分類して保存するために、文書をカテゴリー別に分類する際に使用する分類用辞書を、分類先のカテゴリーが決まっている複数の標本文書を用いて作成し、その後、作成した分類用辞書を用いて分類対象文書をカテゴリー別に分類するということが従来から行われている。

【0003】この種の従来の技術としては、例えば、特開平6-348755号公報に記載されている技術がある。ここでは、各カテゴリーの標本文書内の単語を検出し、唯一のカテゴリーのみに出現した単語をそのカテゴリーのキーワードとして分類用辞書に登録することにより、分類用辞書を作成するようにしている。また、この技術では、分類対象文書中の単語を検出し、更に、分類用辞書に登録済みのキーワードとの一致数を検出し、一致数が最も多かったカテゴリーを分類先とするようにしている。

【0004】また、これ以外にも、例えば、特開平1-188934号公報に記載された技術も従来から知られている。この技術では、複数の標本文書からキーワードを抽出して各キーワードの出現頻度を調べ、更に、カイ二乗計算等の複雑な計算を行って各キーワードの各カテゴリーへの貢献度を示す得点を算出することにより、得

(3)

3

点表（分類用辞書）を作成するようにしている。また、この技術では、分類対象文書からキーワードを抽出し、次いで、抽出した各キーワードに対応する得点を得点表から入力して分類対象文書の各カテゴリ毎の得点を計算し、最も得点の多かったカテゴリを分類対象文書の分類先とするようにしている。

【0005】

【発明が解決しようとする課題】 上述した従来の技術の内、特開平6-348755号公報に記載されている技術では、各カテゴリの標本文書内の単語を検出し、唯一のカテゴリのみに出現した単語をそのカテゴリのキーワードとして分類用辞書に登録することにより、分類用辞書を作成しているため、分類用辞書を作成できない場合があるという問題があった。つまり、特定のカテゴリのみに現れるような単語が存在しない場合は、そのカテゴリについてのキーワードが存在しないことになってしまうため、分類用辞書を作成することができなくなってしまう。

【0006】 また、特開平1-188934号に記載されている技術では、各キーワードの出現頻度に基づいてカイ二乗計算等の複雑な計算を行うことにより分類用辞書を作成しているため、分類用辞書の作成に多くの時間がかかってしまうという問題がある。尚、出現頻度の高いキーワードが必ずしも重要度が高いキーワードであるわけではないので、上述したような複雑な計算を行って分類用辞書を作成するようにしてもあまり意味がない。

【0007】 そこで、本発明の目的は、分類用辞書作成不能となる事態の発生を極力抑え、且つ分類用辞書を短時間で作成できるようにすることにある。

【0008】

【課題を解決するための手段】 本発明の文書分類装置は、上記目的を達成するため、分類用辞書と、標本文書分析結果格納部と、文書の分類先となる複数のカテゴリそれぞれについて、そのカテゴリに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標本文書毎に抽出したキーワードとその標本文書が属するカテゴリのカテゴリ名とを対にして前記標本文書解析結果格納部に格納する標本文書分析部と、前記標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリ名との対に基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対にして前記分類用辞書に格納する代表キーワード抽出部と、前記分類用辞書を使用して分類対象文書をカテゴリ別に分類する文書分類部とを備えている。

【0009】 この構成に於いては、標本文書分析部が、文書の分類先となる複数のカテゴリそれぞれについて、そのカテゴリに属する複数の標本文書を入力して各標本文書に含まれるキーワードを抽出し、更に、各標

4

本文書毎に抽出したキーワードとその標本文書が属するカテゴリのカテゴリ名とを対にして標本文書解析結果格納部に格納し、代表キーワード抽出部が、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリ名との対に基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対にして分類用辞書に格納し、文書分類部が分類用辞書を使用して分類対象文書をカテゴリ別に分類する。

【0010】 また、本発明の文書分類装置は、分類用辞書を作成できなくなるという事態の発生を更に少なくできるようにするため、代表キーワード抽出部の代わりに、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリ名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標本文書の内の、前記代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対にして前記分類用辞書に格納する拡張代表キーワード抽出部を備えている。

【0011】 この構成に於いては、拡張代表キーワード抽出部が、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリ名との対と、ユーザによって設定された代表キーワード抽出基準とに基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標本文書の内の、代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対にして分類用辞書に格納する。

【0012】

【発明の実施の形態】 次に本発明の実施の形態について図面を参照して詳細に説明する。

【0013】 図1は本発明にかかる文書分類装置の一実施例のブロック図であり、ユーザによって標本文書7とそれが属するカテゴリのカテゴリ名8とが入力される標本文書分析部1と、キーワード抽出部2と、標本文書分析結果格納部3と、代表キーワード抽出部4と、フォルダ部5と、ユーザによって分類対象文書9が入力される文書分類部6とを備えている。

【0014】 フォルダ部5は、図2に示すように、カテゴリ名格納部51と、文書格納部52と、代表キーワード格納部53とから構成されている。そして、カテゴリ名格納部51と代表キーワード格納部53とにより分類用辞書が構成される。

【0015】 カテゴリ名格納部51には、文書の分類先とする全てのカテゴリのカテゴリ名が格納される。文書格納部52には、カテゴリ名格納部51に格納されているカテゴリ名が示すカテゴリに属する文

(4)

5

書が格納される。代表キーワード格納部5-3には、カテゴリ名格納部5-1に格納されているカテゴリ名が示すカテゴリを代表する代表キーワードが格納される。尚、初期状態に於いては、フォルダ部5の各部5-1～5-3は、情報が全く格納されていない状態になっている。

【0016】図2の例は、文書の分類先とするカテゴリが「入会手続」、「顧客管理」、「操作方法」、「カタログ請求」、「支払請求」であり、各カテゴリを代表する代表キーワードがそれぞれ「入会、手続、住所、名前」、「住所、電話、変更」、「変更、方法」、「カタログ、住所、送付」、「料金、支払」であることを示している。更に、図2の例は、各カテゴリに属する文書としてそれぞれ「文書1、文書6、文書9」、「文書3、文書7、文書10」、「文書2、文書4、文書12」、「文書8、文書13」、「文書5、文書11、文書14」が格納されていることを示している。

【0017】標本文書分析部1は、ユーザによって標本文書7とカテゴリ名8が入力されると、キーワード抽出部2に標本文書7を渡してキーワードを抽出させる機能、キーワード抽出部2によって抽出されたキーワードとカテゴリ名8とを対にして標本文書分析結果格納部3に格納する機能を有する。

【0018】キーワード抽出部2は、標本文書分析部1から標本文書7が渡された場合は、標本文書7に含まれているキーワードを抽出して標本文書分析部1に返し、文書分類部6から分類対象文書9が渡された場合は、分類対象文書9に含まれているキーワードを抽出して文書分類部6に返す機能を有する。尚、キーワードの抽出方法は、キーワードを抽出できればどのような方法でも良く、例えば、キーワード抽出対象文書に対して形態素解析を行い、それに含まれている名詞をキーワードとして抽出するといった方法や、その他の周知の方法をとることができる。

【0019】代表キーワード抽出部4は、標本文書分析結果格納部3に格納されている各標本文書のキーワードとカテゴリ名との対に基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出する機能、抽出した各カテゴリの代表キーワードとそれが代表するカテゴリのカテゴリ名とを対応付けてフォルダ部5の代表キーワード格納部5-3、カテゴリ名格納部5-1に格納する機能を有する。

【0020】文書分類部6は、ユーザによって分類対象文書9が入力されると、それをキーワード抽出部2に渡して分類対象文書9に含まれるキーワードを抽出させる機能や、フォルダ部5のカテゴリ名格納部5-1、代表キーワード格納部5-3（上述したように、両者によって分類用辞書が構成される）を参照して上記抽出したキーワードと一致する代表キーワードを最も多く含むカテゴリを分類対象文書9の分類先のカテゴリとする機能

6

や、分類対象文書9を分類先のカテゴリのカテゴリ名に対応付けてフォルダ部5に格納する機能を有する。

【0021】図3は標本文書分析部1の処理例を示す流れ図、図4は代表キーワード抽出部4の処理例を示す流れ図、図5は文書分類部6の処理例を示す流れ図であり、以下各図を参照して本実施例の動作を説明する。

【0022】ユーザは、先ず、分類用辞書を作成するために、或るカテゴリに属する標本文書7と、それが属するカテゴリのカテゴリ名8とを標本文書分析部1に入力する。尚、標本文書7は、既に作成済みの文書の中から適切な文書を選び出すようにした方が、作業を効率的に行うことができるが、新たに作成するようにしても構わない。

【0023】標本文書分析部1は、標本文書7、カテゴリ名8が入力されると、図3の流れ図に示すように、標本文書7をキーワード抽出部2に渡す（S31）。これにより、キーワード抽出部2は、標本文書7に含まれている全てのキーワードを抽出して標本文書分析部1に返す。

【0024】キーワード抽出部2からキーワードを受け取ると（S32）、標本文書分析部1は、ユーザによって入力されたカテゴリ名8とキーワード抽出部2から渡されたキーワードとを対にしたレコードを標本文書分析結果格納部3に格納する（S33）。

【0025】今、例えば、ユーザが、カテゴリ名8が「入会手続」の標本文書7を入力したとすると、標本文書分析部1は、先ず、標本文書7をキーワード抽出部2に渡してそれに含まれる全てのキーワードを抽出させる（S31）。キーワード抽出部2で抽出されたキーワードが「入会、手続、住所、料金、電話」であったとすると、標本文書分析部1は、カテゴリ名「入会手続」とキーワード「入会、手続、住所、料金、電話」とを対にしたレコードを標本文書分析結果格納部3に格納する。

【0026】ユーザは、文書の分類先とする全てのカテゴリについて、それぞれ複数の異なる標本文書7を入力すると共にカテゴリ名8を入力する。標本文書分析部1は、ユーザによって標本文書7、カテゴリ名8が入力される毎に上述したと同様の処理を行う。

【0027】図6は、ユーザが全ての標本文書7を入力し終わった時の標本文書分析結果格納部3の内容例を示した図である。この例は、カテゴリ名「入会手続」、「顧客管理」、「操作方法」、…の標本文書がそれぞれ5文書、3文書、3文書、…ずつ入力された時の例である。

【0028】ユーザは、全ての標本文書7を入力すると、代表キーワード抽出部4を動作させる。

【0029】これにより、代表キーワード抽出部4は、図4に示すように、標本文書分析結果格納部3に格納されているレコード中にカテゴリ名の1つに注目する（S41）。その後、代表キーワード抽出部4は、注目

(5)

7

したカテゴリ名を含む全てのレコードを標本文書分析結果格納部3から取り出す(S42)。

【0030】今、例えば、標本文書分析結果格納部3の内容が図6に示すもので、注目しているカテゴリ名が「入会手続」であったとすると、代表キーワード抽出部4は、図7に示す5個のレコードを標本文書分析結果格納部3から取り出すことになる。

【0031】その後、代表キーワード抽出部4は、S42で取り出した全てのレコードに含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対応付けてフォルダ部5の代表キーワード格納部53、カテゴリ名格納部51に格納する(S43、S44)。

【0032】今、例えば、S42で図7に示した5個のレコードを標本文書分析結果格納部3から取り出したとすると、代表キーワード抽出部4は、5個のレコードの全てに含まれている3個のキーワード「入会、手続、住所」を代表キーワードとして抽出し、抽出した代表キーワード「入会、手続、住所」とカテゴリ名「入会手続」とをフォルダ部5の代表キーワード格納部53、カテゴリ名格納部51に格納することになる(S43、S44)。

【0033】代表キーワード抽出部4は、未注目のカテゴリ名がなくなるまで(S45がNOとなるまで)、上記した処理を繰り返し行う。図8は、未注目のカテゴリ名がなくなるまで、上記した処理を行った時にフォルダ部5の内容例を示した図であり、この例は、カテゴリ名「入会手続」、「顧客管理」、「操作方法」、「カタログ請求」、「支払請求」のカテゴリの代表キーワードがそれぞれ「入会、手続、住所」、「住所、電話、変更」、「変更、方法」、「カタログ、住所、送付」、「料金、支払」であることを示している。

【0034】フォルダ部5のカテゴリ名格納部51、代表キーワード格納部53に、上述したようにしてカテゴリ名、代表キーワードを格納し、分類を行うための準備が完了すると、ユーザは、分類対象文書9を文書分類部6に入力する。

【0035】文書分類部6は、分類対象文書9が入力されると、図5の流れ図に示すように、分類対象文書9をキーワード抽出部2に渡す(S51)。これにより、キーワード抽出部2は、分類対象文書9に含まれているキーワードを全て抽出し、抽出したキーワードを全て文書分類部6に返す。

【0036】文書分類部6は、キーワード抽出部2からキーワードを受け取ると(S52)、フォルダ部5のカテゴリ名格納部51、代表キーワード格納部53を参照し、キーワードと一致する代表キーワードを最も多く含むカテゴリ名を求め(S53)、その後、S53で求めたカテゴリ名と対応付けてフォルダ部5の文書格納部52に分類対象文書9を格納する(S54)。

8

【0037】今、例えば、分類対象文書9として、「入会したいので、手続に必要な資料を下記の住所まで送付して下さい。」が入力され、フォルダ部5の内容が図8に示すものであったとすると、以下の処理が行われることになる。

【0038】先ず、文書分類部6が、入力された分類対象文書「入会したいので、手続に必要な資料を下記の住所まで送付して下さい。」をキーワード抽出部2に渡す(S51)。

【0039】これにより、キーワード抽出部2がキーワードを抽出し、抽出したキーワードを文書分類部6に返す。今、例えば、キーワードとして「入会、手続、資料、住所、送付」が抽出されたとする。

【0040】文書分類部6は、キーワード抽出部2から上記したキーワード「入会、手続、資料、住所、送付」を受け取ると(S52)、キーワードと一致する代表キーワードを最も多く含むカテゴリ名を求める(S53)。図8の例では、上記したキーワードと一致する代表キーワードの数は、「入会手続」で3個、「顧客管理」で1個、「操作方法」で0個、「カタログ請求」で2個、「支払請求」で0個となるので、最も一致数の多い「入会手続」が選ばれることになる。

【0041】その後、文書分類部6は、フォルダ部5中の文書格納部52の、カテゴリ名「入会手続」と対応する部分に、入力された分類対象文書「入会したいので、手続に必要な資料を下記の住所まで送付して下さい。」を格納する。

【0042】図9は、本発明にかかる文書分類装置の他の実施例のブロック図であり、図1に示した文書分類装置との相違点は、代表キーワード抽出部4の代わりに拡張代表キーワード抽出部4'を備えた点、代表キーワード抽出基準10を入力する代表キーワード抽出基準入力部11を追加した点である。尚、図9に於いて、他の図1と同一符号は、同一部分を表している。

【0043】代表キーワード抽出基準入力部11は、ユーザが設定した代表キーワード抽出基準10を入力する機能を有する。ここで、代表キーワード抽出基準10は、代表キーワードの抽出基準を示すものであり、 A/B ($A \leq B$) の形式を有する。つまり、或るカテゴリに属する複数の標本文書の内の、代表キーワード抽出基準10によって示される割合 A/B 以上の標本文書に含まれているキーワードを代表キーワードにすることを指示するものである。尚、代表キーワード抽出基準10の値は、変更可能なものである。

【0044】拡張代表キーワード抽出部4'は、標本文書分析結果格納部3に格納されている図6に示すような各レコードと、代表キーワード抽出基準入力部11によって入力された代表キーワード抽出基準10とに基づいて、各カテゴリ毎に、そのカテゴリに属する全ての文書の内の、代表キーワード抽出基準10によって示さ

(6)

9

れる割合を超える標本文書に含まれるキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対応付けてフォルダ部5に格納する機能を有する。

【0045】図10は拡張代表キーワード抽出部4'の処理例を示す流れ図であり、以下各図を参照して本実施例の動作を説明する。

【0046】ユーザは、まず、分類用辞書を作成するために、或るカテゴリに属する標本文書7と、それが属するカテゴリのカテゴリ名8とを標本文書分析部1に入力する。

【0047】標本文書分析部1は、標本文書7、カテゴリ名8が入力される毎に前述したと同様の処理を行い(図3、S31～S33)、標本文書分析結果格納部3に図6に示すようなレコードを格納する。

【0048】ユーザは、分類用辞書を作成するために用意した全ての標本文書7を入力すると、代表キーワード抽出基準10を設定する。これにより、代表キーワード抽出基準入力部11が、代表キーワード抽出基準10を拡張代表キーワード抽出部4'に入力する。

【0049】代表キーワード抽出基準10が入力されると、拡張代表キーワード抽出部4'は、図10の流れ図に示すように、標本文書分析結果格納部3に格納されているレコードに含まれているカテゴリ名の内の1つに注目する(S101)。

【0050】その後、拡張代表キーワード抽出部4'は、注目したカテゴリ名を含む全てのレコードを標本文書分析結果格納部3から取り出す(S102)。

【0051】次いで、拡張代表キーワード抽出部4'は、S102で取り出したレコードの内の、代表キーワード抽出基準10によって示される割合以上のレコードに含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対応付けてフォルダ部5の代表キーワード格納部53、カテゴリ名格納部51に格納する(S103、S104)。その後、拡張代表キーワード抽出部4'は、未注目のレコードがなくなるまで(S105がNOとなるまで)、上述した処理を繰り返し行う。

【0052】今、例えば、標本文書分析結果格納部3の内容が図6に示すものであり、ユーザが設定した代表キーワード抽出基準10が「3/5」であったとすると、以下の処理が行われる。

【0053】まず、標本文書分析結果格納部3に格納されているカテゴリ名の内の1つ(「入会手続」とする)に注目する(S101)。次に、標本文書分析結果格納部3からカテゴリ名が「入会手続」となっているレコードを全て取り出す(S102)。この例の場合、図7に示す5個のレコードが取り出される。

【0054】その後、図7に示した5個のレコードの内の、代表キーワード抽出基準10によって示される割合

10

「3/5」以上のレコードに含まれるキーワードを代表キーワードとして抽出する(S103)。この例の場合、キーワードが含まれる割合は、「入会」=5/5、「手続」=5/5、「住所」=5/5、「料金」=4/5、「名前」=4/5、「電話」=2/5、「FAX」=1/5、「方法」=1/5であるので、拡張代表キーワード抽出部4'は、「入会」、「手続」、「住所」、「料金」、「名前」の5つを代表キーワードとして抽出する。

【0055】その後、拡張代表キーワード抽出部4'は、カテゴリ名「入会手続」と代表キーワード「入会、手続、住所、料金、名前」とを対応付けてフォルダ部5に格納する。

【0056】以下、拡張代表キーワード抽出部4'は、未注目のレコードがなくなるまで(S105がNOとなるまで)、上述した処理を繰り返し行う。図11は、未注目のレコードがなくなるまで、上述した処理を行った時のフォルダ部5の内容を示した図である。

【0057】フォルダ部5のカテゴリ名格納部51、代表キーワード格納部53に、上述したようにしてカテゴリ名、代表キーワードを格納し、分類を行うための準備が完了すると、ユーザは、分類対象文書9を文書分類部6に入力する。

【0058】文書分類部6は、分類対象文書9が入力されると、前述したと同様にして、分類対象文書9を文書格納部52に該当する部分に格納する。

【0059】図12は、図1及び図9に示した文書分類装置のハードウェア構成を示したブロック図であり、コンピュータ121と、記録媒体122と、記憶装置123とを備えている。記録媒体122は、磁気ディスク、半導体メモリ、その他の記録媒体である。

【0060】図1に示した文書分類装置を実現する場合には、記録媒体122に記録された文書分類用プログラムがコンピュータ121によって読み込まれ、コンピュータ121の動作を制御することで、コンピュータ121上に図1に示した標本文書分析部1、キーワード抽出部2、代表キーワード抽出部4、文書分類部6を実現する。尚、標本文書分析結果格納部3、フォルダ部5は、記憶装置123上に構成される。

【0061】また、図9に示した文書分類装置を実現する場合には、記録媒体122に記録された文書分類用プログラムがコンピュータ121によって読み込まれ、コンピュータ121の動作を制御することで、コンピュータ121上に図9に示した標本文書分析部1、キーワード抽出部2、拡張代表キーワード抽出部4'、文書分類部6、代表キーワード抽出基準入力部11を実現する。尚、標本文書分析結果格納部3、フォルダ部5は、記憶装置123上に構成される。

【0062】

【発明の効果】以上説明したように、本発明の文書分類

(7)

11

装置は、標本文書分析結果格納部に格納されている各標本文書のキーワードとカテゴリ名との対に基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対にして分類用辞書に格納することにより分類用辞書を作成する代表キーワード抽出部を備えているので、唯一のカテゴリのみに出現した単語をそのカテゴリのキーワードとして分類用辞書に登録することにより分類用辞書を作成するようにしていた従来技術に比較して、分類用辞書が作成不能となる危険性を少なくすることができ、且つカイ二乗計算等の複雑な計算を行うことにより分類用辞書を作成していた従来に技術に比較して分類用辞書の作成時間を短くすることができる。

【0063】また、本発明の文書分類装置は、ユーザによって指定された代表キーワード抽出基準と、標本文書解析結果格納部の格納されている各標本文書のキーワードとカテゴリ名との対に基づいて、各カテゴリ毎に、そのカテゴリに属する全ての標準文書の内の、代表キーワード抽出基準によって示される割合以上の標本文書に含まれているキーワードを代表キーワードとして抽出し、抽出した代表キーワードとカテゴリ名とを対にして分類用辞書に格納することにより分類用辞書を作成するようにしているので、分類用辞書が作成不能となる危険性を更に少なくすることができる。更に、或るカテゴリを示す重要なキーワードが全ての標本文書に含まれていなかったとしても、そのキーワードを代表キーワードとして含む分類用辞書を作成することができる。

【図面の簡単な説明】

【図1】本発明の一実施例のブロック図である。

【図2】フォルダ部5の内容例を示す図である。

【図3】標本文書分析部1の処理例を示す流れ図である。

【図4】代表キーワード抽出部4の処理例を示す流れ図である。

【図5】文書分類部6の処理例を示す流れ図である。

【図7】

カテゴリ名	キーワード
入会手続	入会、手続、住所、料金、電話
入会手続	入会、手続、住所、名前、料金、電話
入会手続	入会、手続、住所、名前、料金、FAX
入会手続	入会、手続、住所、名前、料金
入会手続	入会、手続、住所、名前、方法

12

【図6】標本文書分析結果格納部3の内容例を示す図である。

【図7】代表キーワード抽出部4が標本文書分析結果格納部3から取り出した、カテゴリ名が同一のレコードを示した図である。

【図8】カテゴリ名格納部51、代表キーワード格納部53にカテゴリ名、代表キーワードを格納し終わった時のフォルダ部5の内容例を示す図である。

【図9】本発明の他の実施例のブロック図である。

【図10】拡張代表キーワード抽出部4'の処理例を示す流れ図である。

【図11】カテゴリ名格納部51、代表キーワード格納部53にカテゴリ名、代表キーワードを格納し終わった時のフォルダ部5の内容例を示す図である。

【図12】文書分類装置のハードウェア構成の一例を示すブロック図である。

【符号の説明】

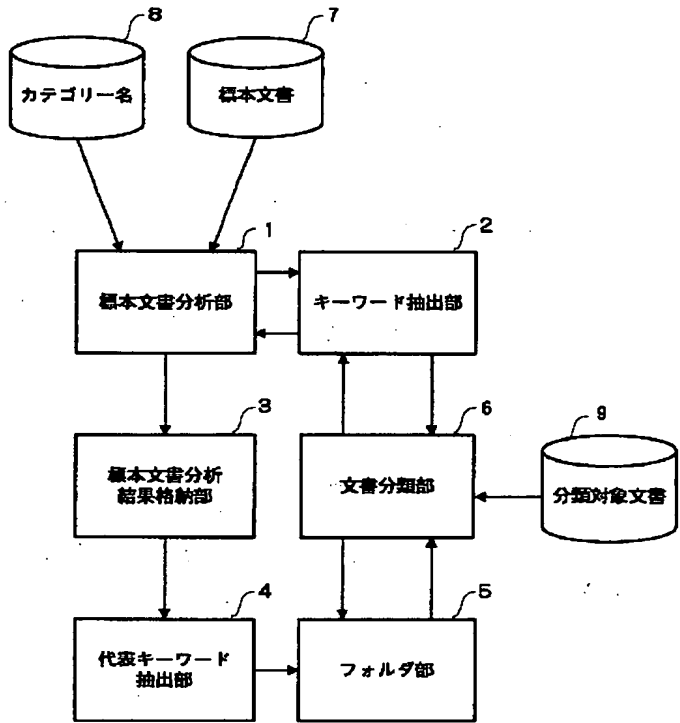
- 1…標本文書分析部
- 2…キーワード抽出部
- 3…標本文書分析結果格納部
- 4…代表キーワード抽出部
- 4'…拡張代表キーワード抽出部
- 5…フォルダ部
- 51…カテゴリ名格納部
- 52…文書格納部
- 53…代表キーワード格納部
- 6…文書分類部
- 7…標本文書
- 8…カテゴリ名
- 9…分類対象文書
- 10…代表キーワード抽出基準
- 11…代表キーワード抽出基準入力部
- 121…コンピュータ
- 122…記録媒体
- 123…記憶装置

【図8】

カテゴリ名	カテゴリに属する文書	代表キーワード
入会手続		入会、手続、住所
顧客管理		住所、電話、変更
操作方法		変更、方法
カタログ請求		カタログ、住所、送付
支払請求		料金、支払

(8)

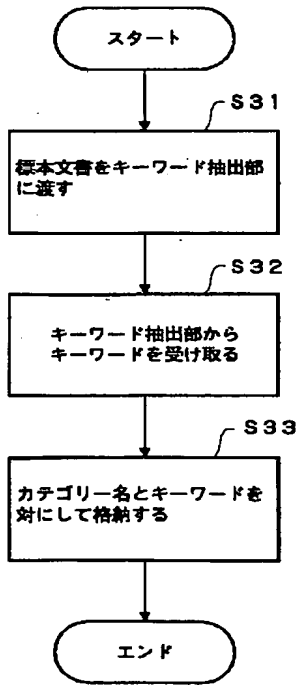
【図1】



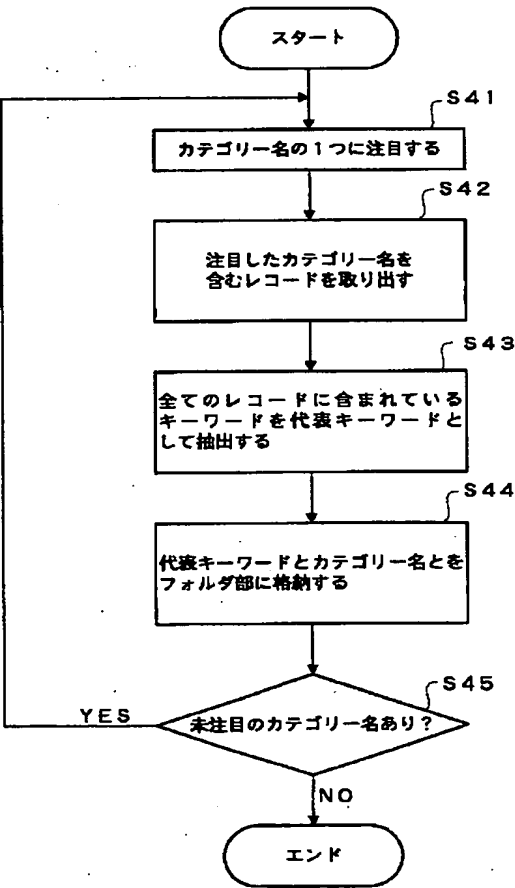
【図2】

51 カテゴリ名	52 カテゴリに属する文書	53 代表キーワード
入会手続	文書1, 文書8, 文書9	入会, 手続, 住所, 名前
顧客管理	文書3, 文書7, 文書10	住所, 電話, 変更
操作方法	文書2, 文書4, 文書12	変更, 方法
カタログ請求	文書8, 文書13	カタログ, 住所, 送付
支払請求	文書5, 文書11, 文書14	料金, 支払

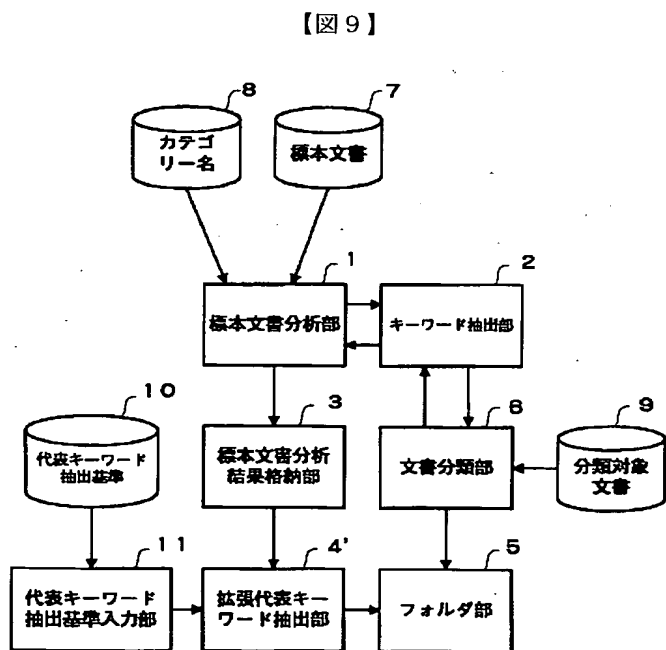
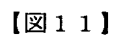
【図3】



【図4】



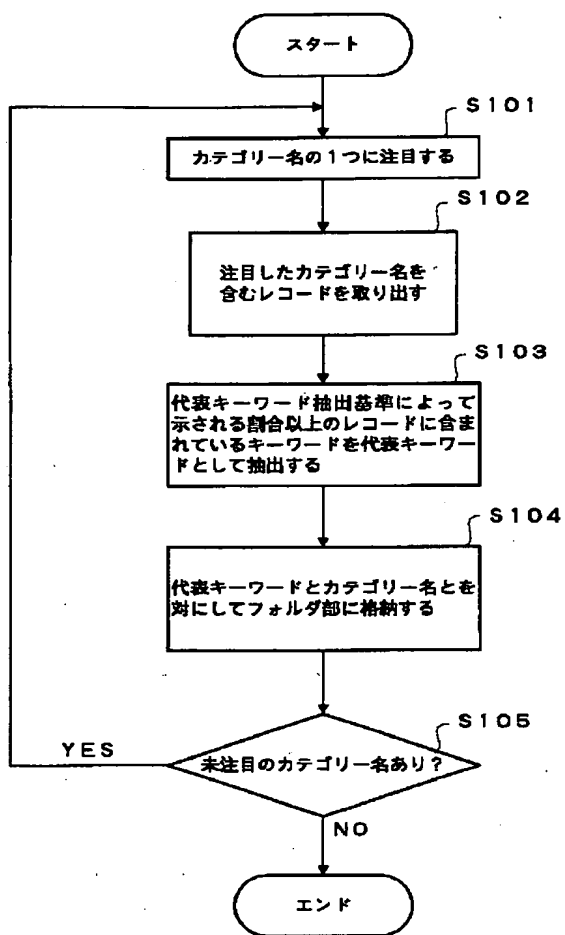
【図 6】



51	52	53
カテゴリ名	カテゴリに属する文書	代表キーワード
入会手続		入会, 手続, 住所, 料金, 名前
顧客管理		住所, 電話, 変更
操作方法		変更, 方法
カタログ請求		カタログ, 住所, 送付
支払請求		料金, 支払

(10)

【図10】



【図12】

